J.D. Lohn, S.P. Colombano, J. Scargle, D. Stassinopoulos, G.L. Haith, "Evolving Catalytic Reaction Sets using Genetic Algorithms," *1998 IEEE International Conference on Evolutionary Computation*, 1998, to appear.

Evolving Catalytic Reaction Sets using Genetic Algorithms

Jason D. Lohn, Silvano P. Colombano, Jeffrey Scargle, Dimitris Stassinopoulos, Gary L. Haith

Abstract— **In this paper we construct simple artificial chemistries in order to gain an understanding of how a chemical reaction network might emerge from a state of relative disorder in non-living "protocells." Such chemistries have relevance to origin of life studies as well as artificial life research. We present a model comprised of interacting polymers, and specify two initial conditions: a distribution of relatively disordered polymers and a fixed set of reversible catalytic reactions. A genetic algorithm is then used to find a set of reactions that exhibit pre-specified behavior. Our results show that reaction sets can be found to give polymer distributions that are biased towards longer polymers. We present examples of these protocell chemistries and show that the reaction sets found are robust in the sense that they produce desirable behavior in equilibrium. Such a technique is useful because it allows an investigator to determine whether a specific distribution can be produced, and if it can, a reaction network can be found and then analyzed.**

I. Introduction

Protocells are conjectured to be the precursors to the first living cells on Earth. As membrane-bound structures, protocells may have been capable of basic functions which utilize the simple molecules thought to exist under prebiotic conditions [9]. Protocells could have formed by the self-assembly of bilayer membranes. Studies have found that amphiphilic molecules spontaneously accumulate at water/air and water/oil interfaces, and self-assemble into membrane structures by agitation or cycles of wetting and drying [7].

Given the formation of protocells, the pathway to the emergence of cellular life would likely involve metabolism formation and reproduction. Our focus here is in the development of primitive protocellular metabolisms subsequent to membrane formation. The emergence of catalytic and autocatalytic reaction sets is thought to be a crucial step in the evolution of metabolisms [6]. We simulate a single protocell containing short polymers and construct an artificial chemistry to study how a chemical reaction network might emerge and organize beginning from a state of relative disorder. Recent work in this direction includes the autocatalytic reaction networks of Bagley, et. al. [1], and the work of Fontana and Buss where a λ-calculus formalism is employed to construct artificial chemistries [4]. Also relevant are Schuster's models of selection for autocatalytic reactions (e.g. [10]).

J. Lohn is with Caelum Research Corporation at NASA Ames Research Center, MS 269/1, Moffett Field, CA, 94035.

S. Colombano and D. Stassinopoulos are with the Computational Sciences Division of NASA Ames Research Center, MS 269/1, Moffett Field, CA, 94035.

J. Scargle is with the Planetary Systems Branch of NASA Ames Research Center, MS 245/3, Moffett Field, CA, 94035.

G. Haith is with the Psychology Department, Stanford University, Stanford, CA 94305-2130.

II. Protocell Model

The protocell model we use is a simple mass-conserving, well-stirred reactor. The simulated molecules are linear polymers (chains) comprised of a single type of molecule, a. The two types of interaction are bonding (condensation) and breaking (cleavage) of simulated chemical bonds. Since our polymers are chains of a single type, bonding and breaking function like arithmetic addition and subtraction: bonding a monomer a to a 4-mer $aaaa$ results in a 5-mer, $aaaaa$. Bonding and breaking operations are influenced by efficiencies – parameters which reflect the probability that a given polymer will undergo a specific operation. For the results reported here, we used a small protocell reactor in which we limit polymer length to 34. This in turn limits the number of allowable reactions, however the space of all possible reaction sets is still quite large (on the order of 10^{400} for a reaction set containing 100 reactions).

A fixed initial distribution of polymers is used in all experiments, and is of the form of a decaying exponential: $f(x) = 100e^{-(x-1)/10}$. Such a distribution is a relatively disordered initial state (many short and few long polymers), and is a plausible initial condition for a protocell since larger polymers are more likely to break as compared to smaller ones. We assume all polymers are enclosed by an impermeable membrane, and polymer interactions occur at random (see Figure 1). We arbitrarily restrict interactions to be catalytic, as discussed in the next section.

Time is simulated using "reaction cycles," periods in which the same number of polymer interactions occur, which is generally different than the number of reactions executed. For example, if two polymers are

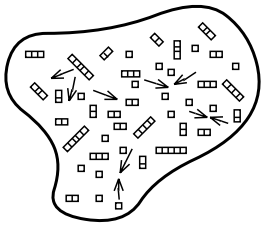


Fig. 1. Interacting polymers enclosed in a protocell.

chosen from the reactor (e.g., a 5-mer and 24-mer), and there does not exist a reaction in the reaction list that binds them together (to form a 29-mer), then no reaction takes place, but it counts as an interaction.

III. Catalytic Reaction Sets

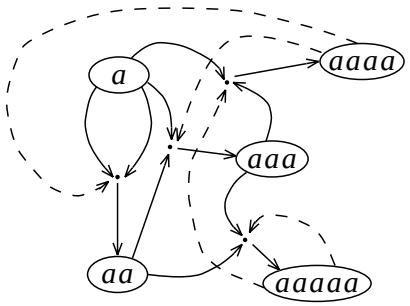
Given polymers A, B, C, and P, the reactions we use in our model are of the following form:

$$A + B \overset{C}{\rightleftharpoons} P$$

where A, B, and P are reactants, and C is a catalyst. Reactions are reversible, so that a product can break into two shorter polymers when one of its bonds is broken. Catalysts are polymers which increase the speed of chemical reactions. In our model, since we do not take into account reaction rates, the catalysts act as switches: if the catalyst is present, a reaction can proceed, otherwise the reaction cannot. Also, our catalysts do not undergo any changes as can happen in real reactions. Upon completion of a reaction, the catalyst remains and may further catalyze other reactions. An autocatalytic reaction occurs when one of the reactants also catalyzes the reaction. Such reactions are permitted in our model.

We denote the set of all possible catalytic reactions, whether biochemically realistic or not, as R. We then have two subsets of interest. Let R_b be the subset of reactions that are biochemically realistic, and R_n be a subset containing at most n reactions from R. For our current study, we will use R_n with $n = 100$ in our model and therefore in the genetic algorithm (GA) search. The set R_b is the clear choice for modeling prebiotic chemistries on Earth, and we plan to use R_b in future models.

Because of the interrelationships that can form within a catalytic set, we can view them as graph structures. A simple catalytic reaction set is depicted in Figure 2, where four reactions are shown. The symbol "$\cdot$" denotes a reaction, and dashed lines denote catalysts. Since all reactions are catalyzed by polymers within the network, this is an example of an autocatalytic network. As an autocatalytic network, this example shows how such networks can be thought of as small chemical "engines." Given an abundant supply of monomers, and a high forward reaction efficiency, we can readily see the main output of this engine: production of 4-mers and 5-mers.



$$a + a \overset{aaaa}{\rightleftharpoons} aa \qquad a + aaa \overset{aaaa}{\rightleftharpoons} aaaa$$

$$a + aa \overset{aaaa}{\rightleftharpoons} aaa \qquad aa + aaa \overset{aaaaa}{\rightleftharpoons} aaaaa$$

Fig. 2. Example of a catalytic reaction set: graphical depiction (top) and reaction set (bottom). Reverse reactions are not shown explicitly in graph.

Reaction networks in the real world can exhibit far greater complexity, and can be seen as adaptive systems that exhibit interesting dynamical behaviors. Representing the chemical reactions as a reaction graph allows easier identification of cycles, dependencies, and other properties.

IV. Experimental Setup

To automatically produce reaction sets we used a genetic algorithm [5] as our search technique. Our interest here is not to compare search techniques, although that would be interesting and could result in more effective search. Rather, since we are satisfied with approximate solutions, and we anticipate scaling up our simulations to handle vastly more complex polymers with differing initial conditions (resulting in extremely large search spaces), we find the flexibility and effectiveness of the genetic algorithm well-suited for these purposes.

Our objective is finding reaction sets (or equivalently, reaction graphs) that can take the simple protocell polymers from an initial "disordered" distribution (of polymer lengths) to one that is biased towards building up long polymers. We choose two target polymer distributions, called `peak` and `target`, that reflect this buildup. These are shown in Figures 3 and 4, where each shows the target along with the exponentially decreasing initial distribution. The

first is a flat peak of 70 polymers for the polymers of lengths 21, 22, 23. The second is a linear increase (slope of 10) for polymers of lengths 25–30. The choice of these specific distributions was arbitrary, the main requirements being an emphasis on forming long polymers. Note that the targets are specified for a small number of polymer lengths, with the remaining lengths designated as don't care.

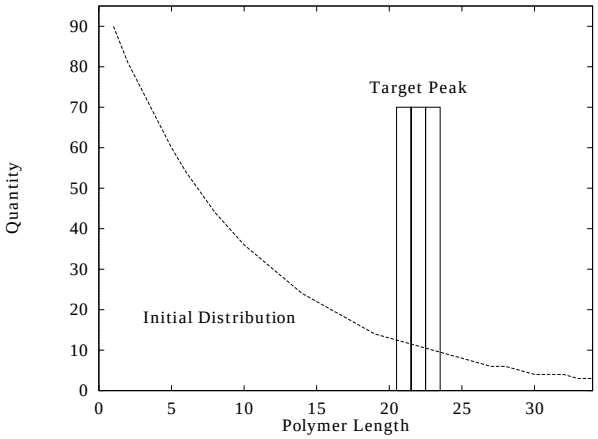


Fig. 3. The `peak` target distribution.

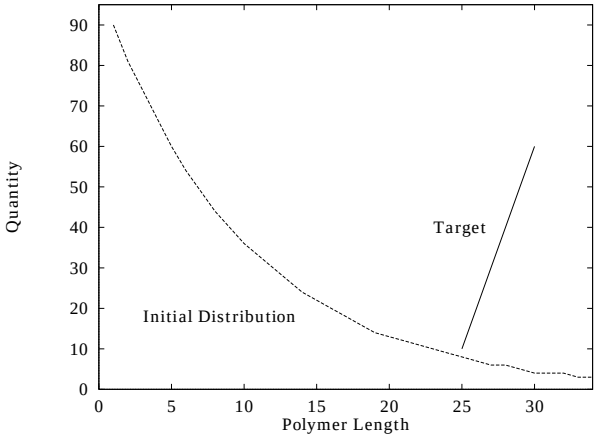


Fig. 4. The `slope` target distribution.

Reaction sets were represented in the GA as arrays of reactions, with each element of the array containing the reactants A, B, C, as well as reaction efficiencies (both forward and reverse). Note that explicit encoding for the reaction product P is not needed, and reverse reactions are implicitly encoded using this method. Because we fixed the maximum length of polymers to 34, we needed to prevent reactions that formed 35-mers and above from occurring. To accomplish this, A is constrained to represent polymers of length 1 through 33, and B is constrained to represent polymers of length $(34 - \text{len}(A))$. In this way there are 289 combinations of A and B (from $\lceil (N+1)(N-1)/4 \rceil$, with $N = 34$). Each reaction set contains 100 reactions (200 if forward and reverse reactions are counted separately), thus each GA individual is an array of 100 elements. Excluding reaction efficiencies, each reaction instance is one out of $289 \cdot 34 = 9826$ possible reactions, and thus each reaction set instance is one out of slightly less than 9826^{100} possible reaction sets.

The fitness of a given reaction set is computed as follows. At the end of each reaction cycle, an error is computed. The error is the absolute value between the target and simulation quantity summed over all fitness cases. The overall fitness is the minimum error seen over all reaction cycles. Therefore a fitness of zero is a perfect fit, and high-valued fitness values reflect poor fitting of the target distribution (i.e., our fitness function is a cost function, one that we wish to minimize). Each reaction cycle allowed for 100 polymer interactions, and the total number of reaction cycles was limited to 50. The GA is run for a maximum of 200 generations using a population size of 500 individuals. The result of the run is designated as the individual having the lowest fitness (error) value.

V. Experimental Results

We ran two experiments, each identical except for the target polymer distributions (see Figures 3 and 4). In each experiment we used the protocell model described above, and ran 100 GA runs for sampling purposes. Each run within an experiment was identical except for the stream of pseudo-random numbers used.

In order to get a sense of how difficult it was to find reaction sets that promote increasing complexity, we plot the distribution of best-individual fitnesses from each GA run in Figure 5. As can be seen, an error value of five was the most frequent result in the `peak` experiment (mean fitness 7.32) and an error value of 18 was most frequently found in the `slope` experiment (mean fitness 18.09). Only one run (in the `peak` experiment) produced an individual with perfect fitness (error of zero). From these plots we can confirm that the `slope` problem is more difficult – since the `slope` problem has more points to fit. Also, we can say that the lowest error reaction sets are not anomalous – the results are all clustered near low values of error.

The results from the `peak` experiment are as follows. Figure 6 shows the distribution of polymers resulting from the best individual from generation 0. We see that this randomly-generated reaction set is able to

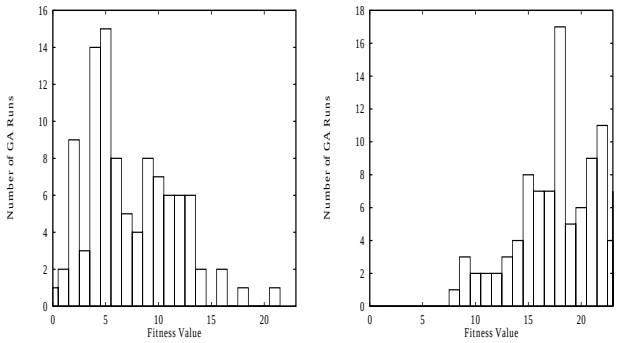


Fig. 5. Histogram showing frequency of best fitness values found in 100 genetic algorithm runs. The left graph is for the `peak` experiment and the `slope` experiment is shown on the right. Lower fitness values correspond to lower error and improved fitness.

effectively convert nearly all of its polymers shorter than length 10 to longer polymers. It also produces approximately half of the required 22-mer and 23-mer polymers.

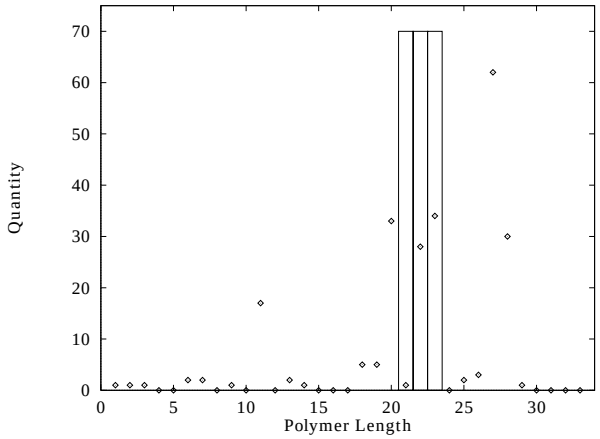


Fig. 6. Distribution resulting from best individual of generation 0 of the `peak` experiment (distribution shown as points, target as peak).

The distribution of polymers resulting from the best individual from the `peak` experiment is shown in Figure 7. The corresponding reaction graph is shown in Figure 8 where we have drawn a portion of the reaction network showing only the most frequently "executed" reactions. The distribution of polymers achieves an error of zero since it exactly matches the target peak. Notable here is the relative abundance of 6-, 7-, 32- and 33-mers. To get a sense as to how the target peak is reached, and why the levels of certain other polymers are elevated, we can look to the reaction graph. There we see that 6-mers, which are in relative abundance in the initial distribution, are catalyzing the production of 34-mers from 33-mers (also elevated), which produce 14-mers by splitting, and 14-mers contribute to 23-mer production (one of the target polymers) via 15-mer production. 6-mers are also important in producing 8-mers which combine with 15-mers to produce more 23-mers. Thus we can see how 6-mers are an integral link in the production chain of 23-mers. 7-mers are also in relative abundance in the initial distribution, however their supply is replenished by 29-mer splits, which also produce 22-mers (another target polymer). Length 7 polymers also contribute to 23-mer production by acting as a catalyst for 15-mer and 8-mer reactions. The remaining target polymer, the 21-mer, is produced by combining 19- and 2-mers, and 17- and 4-mers. 21-mers and 32-mers both catalyze the production of the target 22-mers.

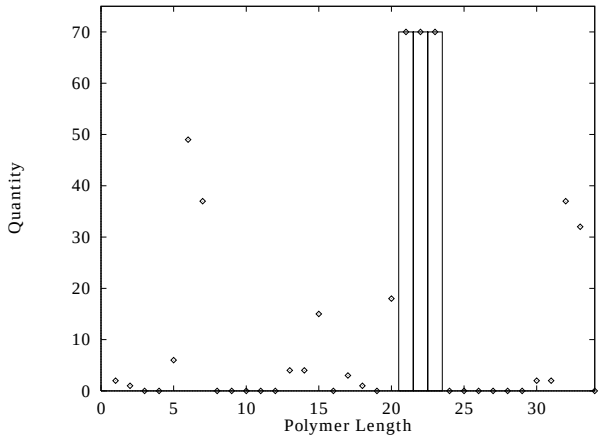


Fig. 7. Best found distribution of polymers (points) and target (peak) for the `peak` experiment.

The reactions in our protocell model can be viewed in a producer/consumer manner. If the reaction $A + B \overset{C}{\rightleftharpoons} P$ proceeds in the forward direction, then A and B are consumed, and P is produced. In the reverse direction, the opposite occurs. In order to build longer polymers, it would follow that the evolved reaction sets would need to be biased to producing longer polymers. This trend can be seen in Figure 9, where we show the distribution of polymers produced and consumed in the best reaction set of the `peak` experiment. With the exception of monomer production, the "producer" distribution (top plot) shows a slight tendency towards longer polymers. In the "consumer" distribution, the bias is more pronounced. There we see a lager bias towards consuming small polymers in order to build larger ones.

The `slope` distribution was more difficult to attain as compared to the `peak` problem. The best found re-

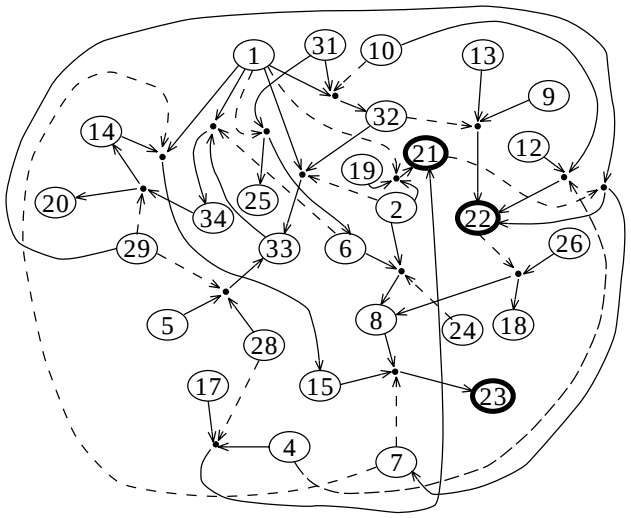


Fig. 8. Reaction graph for highest fitness catalytic set found in the `peak` experiment. Only the most frequently used reactions are shown. Thick ovals indicate the polymers that comprised the target peak.

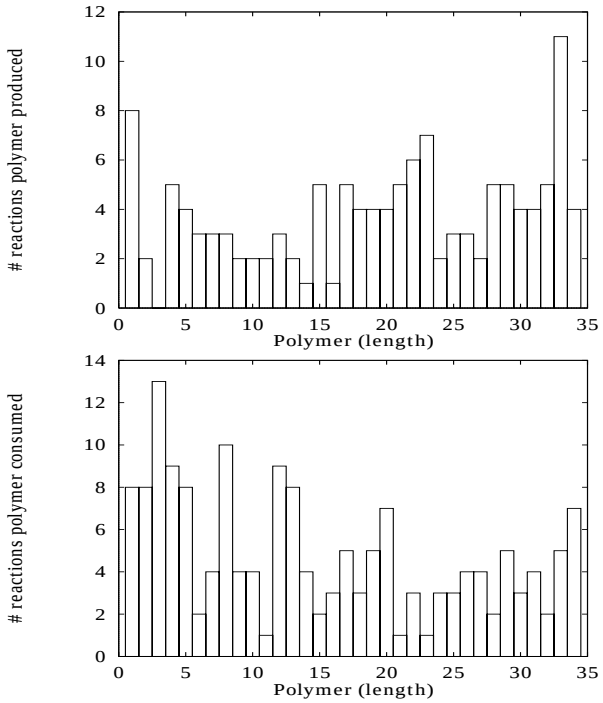


Fig. 9. Distribution of polymer production and consumption in the best-found reaction set for the `peak` experiment. The top plot shows the net number of times each polymer was a product of a reaction, and the bottom plot shows the net number of times each polymer was consumed in reactions.

action set had an error of 8 compared to the zero error found in the `peak` experiment. Figures 10 and 11 show the distributions of polymers at different generations (0 and 118) from the GA run producing the highest fitness reaction set. We can see that the generation zero individual is producing larger polymers, and is able to produce 26-mers in nearly the correct amount. The best individual is able to fit the slope closely by first producing polymers larger than 30, and then breaking those down (via reverse reactions) into products containing 25- through 30-mers. The reaction graph for this reaction set shows many of the same characteristics (short cycle formation, key polymers acting as both reactants and catalysts, target polymers acting as catalysts) as seen in the `peak` reaction graph.

Recall that the fitness evaluation gives feedback to the smallest error seen at a given point in time. Naturally we are also interested in the steady-state, or equilibrium behavior. Although not explicitly designed into the fitness calculation, we find that the resulting reaction sets settle to fixed distributions that are exaggerations of the target distributions. For example, at equilibrium in the `peak` problem, the distribution contains a single large peak for 21-mers only. In the `slope` problem, 27-mers are produced as the sole peak. Thus we can infer that although the evolved catalytic networks have a few strong producer/consumer cycles, each has one in particular that dominates at equilibrium.

We also examined the robustness of the best-found reactions sets by varying the initial distributions. If a given reaction set produces a target peak given the initial distribution it was trained on, it would be interesting to know how sensitive the reaction set is when starting from different initial distributions. Using the best-found reaction sets from the `peak` experiment, we tried a variety of differing initial distributions. In all cases, the fitness values were between 14 and 18. The equilibrium distributions, however, were nearly identical to the evolved equilibrium distributions. This provides evidence that the evolved reaction sets are robust in achieving the target distributions in equilibrium. Further experiments will examine robustness in greater detail.

VI. Conclusion

We have presented highly simplified models of molecular interactions occurring in protocells. We demonstrated that small, artificial chemical reaction networks can be synthesized to move a system of polymers into states of increasing complexity. The reaction sets found are robust in the sense that they produce desirable behavior in equilibrium. Such evolution of complexity is likely fundamental in the development of artificial life structures. And while these

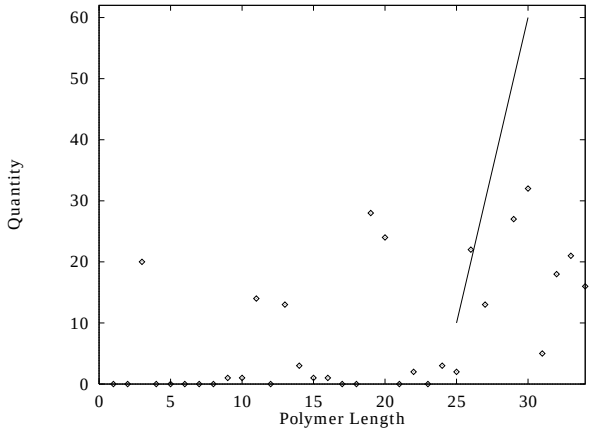


Fig. 10. Distribution resulting from best individual of generation 0 of the `slope` experiment (distribution shown as points, target as line segment).

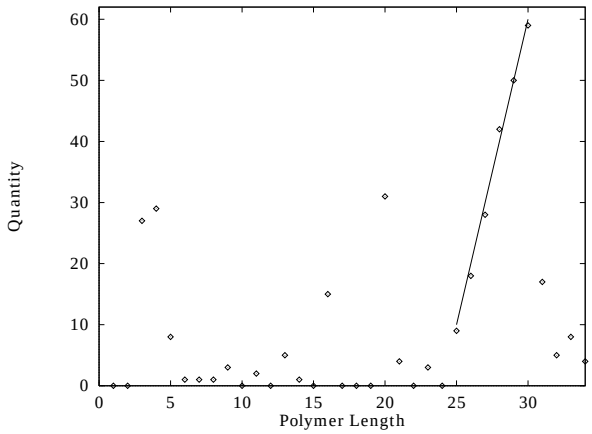


Fig. 11. Best found distribution of polymers (points) and target (line segment) for the `slope` experiment.

chemistries are artificial, they nonetheless point the way toward studies in which chemical realism can be increased. Of great interest is to expand our model of artificial protocellular chemistries to include membrane functions, polymers containing more than one type of molecule, and communities of interacting protocells.

VII. Acknowledgments

The authors thank Bob Hogan, Mike New, Andrew Pohorille, and the anonymous reviewers for their helpful comments and suggestions. D. Stassinopoulos acknowledges support from a National Research Council - NASA Ames Research Associateship.

References

[1] R.J. Bagley, J.D. Farmer, S.A. Kauffman, N.H. Packard, A.S. Perelson, I.M. Stadnyk, "Modeling Adaptive Biological Systems," *BioSystems*, vol. 23, 1989, pp. 113–138.

[2] W. Banzhaf, P. Dittrich, H. Rauhe, "Emergent Computation by Catalytic Reactions," *Nanotechnology*, vol. 7, pp. 307–314, 1996.

[3] J.D. Farmer, S.A. Kauffman, "Autocatalytic Replication of Polymers," *Physica D*, vol. 22, 1986, pgs. 50–67.

[4] W. Fontana, L.W. Buss, "The Arrival of the Fittest: Toward a Theory of Biological Organization," *Bull. Math. Biol.*, vol. 56, 1994, pp. 1-64.

[5] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, Mass, 1989.

[6] S.A. Kauffman, *The Origins of Order*, Oxford University Press, 1993.

[7] D.D. Lasic, *Liposomes: From Physics to Applications*, Elsevier, New York, 1993.

[8] S.L. Miller, L.E. Orgel, *The Origins of Life on Earth*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[9] A. Pohorille, C. Chipot, M.H. New, M.A. Wilson, "Molecular Modeling of Protocellular Functions," *Biocomputing: Proceedings of the 1996 Pacific Symposium*, L. Hunter, T. Klein (Eds.), World Scientific, Singapore, 1996.

[10] P. Schuster, "Evolution of Self-Replicating Molecules – A Comparison of Various Models for Selection," in *Dynamical Systems and Cellular Automata*, J. Demongeot, E. Goles, M. Tchuente (Eds.), Academic Press, 1985, pp. 255–267.